Reverse-Bayes Methods for the Analysis of Replication Studies

Leonhard Held

@HeldLeonhard

University of Zurich UZH

Swiss National Science Foundation

Joint work with Charlotte Micheloud, Samuel Pawel and Fadoua Balabdaoui

O'Bayes 2022

University of California Santa Cruz, US

September 6, 2022

Replicability

- Replicability of research findings is crucial to the credibility of science.
- Large-scale replication projects have been conducted in the last years.
- Such efforts help to assess to what extent results from original studies can be confirmed in independent replication studies.

UZH

Center for Reproducible Science

The Replicability of Psychological Science

Open Science Collaboration, 2015, *Science*


RESEARCH ARTICLE SUMMARY

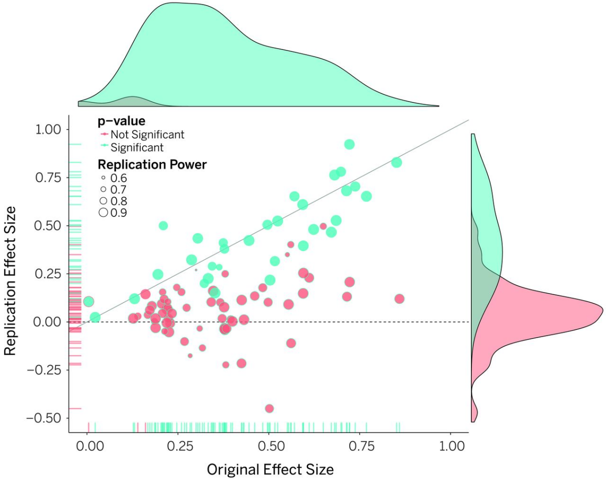
PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

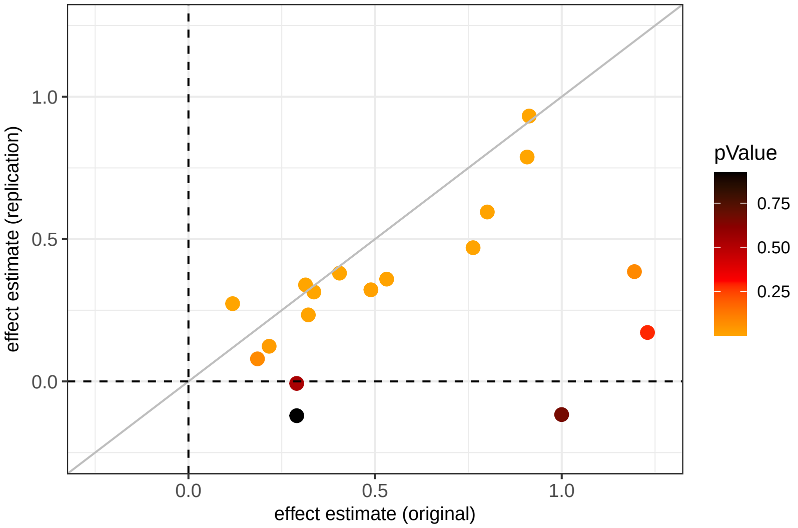
Similar replication projects:

- Experimental Economics (2016)
- Social Sciences (2018)
- Experimental Philosophy (2018)
- Cancer Biology (2021)



Experimental Economics Replication Project

Camerer *et al.* (2016), Science



Replication is Standard in Drug Regulation

- FDA/EMA requires

  “at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness.”

Replication is Standard in Drug Regulation

- FDA/EMA requires

  "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."

- Usually implemented requiring one-sided $p < \alpha = 0.025$ in two independent studies ("two-trials rule").

Replication is Standard in Drug Regulation

- FDA/EMA requires

  "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."

- Usually implemented requiring one-sided $p < \alpha = 0.025$ in two independent studies ("two-trials rule").

$\rightarrow$ Type-I error (T1E) rate is $\alpha^2 = 0.025^2 = 0.000625$

Replication is Standard in Drug Regulation

- FDA/EMA requires

  > *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."*

- Usually implemented requiring one-sided $p < \alpha = 0.025$ in two independent studies ("two-trials rule").

$\rightarrow$ Type-I error (T1E) rate is $\alpha^2 = 0.025^2 = 0.000625$

- However, "double dichotomisation" may not reflect the available evidence:
  - $p_1 = p_2 = 0.024$ leads to claim of success.

Replication is Standard in Drug Regulation

- FDA/EMA requires

  "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."

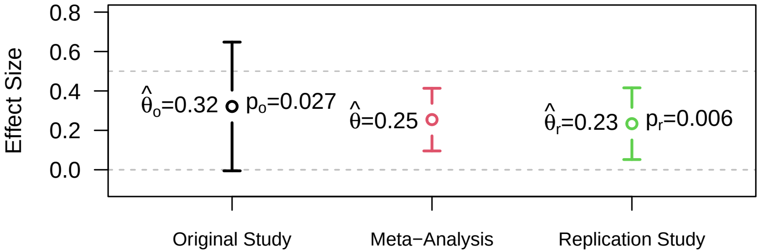
- Usually implemented requiring one-sided $p < \alpha = 0.025$ in two independent studies ("two-trials rule").

$\rightarrow$ Type-I error (T1E) rate is $\alpha^2 = 0.025^2 = 0.000625$

- However, "double dichotomisation" may not reflect the available evidence:
  - $p_1 = p_2 = 0.024$ leads to claim of success.
  - $p_1 = 0.027$ and $p_2 = 0.006$ leads to no claim of success.

Example: Ambrus and Greiner (2012), Experimental Economics

Effect estimates with 95% confidence interval



1. Two-trials rule (one-sided)
2. Compatibility of effect estimates (Q-test): $p_Q = 0.65$
3. Meta-analysis of effect estimates (95% CI): [0.10, 0.41]

Assessement of Replication Success

- Limitations of currently used methods:
  - Two-trials rule is based on “double dichotomisation”
  - Q-test provides no information about true effect
  - Meta-analysis assumes exchangeability

Assessement of Replication Success

- Limitations of currently used methods:
  - Two-trials rule is based on “double dichotomisation”
  - Q-test provides no information about true effect
  - Meta-analysis assumes exchangeability
- I will describe two reverse-Bayes approaches
  - without “double dichotomisation”
  - without exchangeability assumptions
  - but with explicit penalisation of effect size shrinkage

Assesement of Replication Success

- Limitations of currently used methods:
  - Two-trials rule is based on “double dichotomisation”
  - Q-test provides no information about true effect
  - Meta-analysis assumes exchangeability
- I will describe two reverse-Bayes approaches
  - without “double dichotomisation”
  - without exchangeability assumptions
  - but with explicit penalisation of effect size shrinkage

1. The sceptical *p*-value

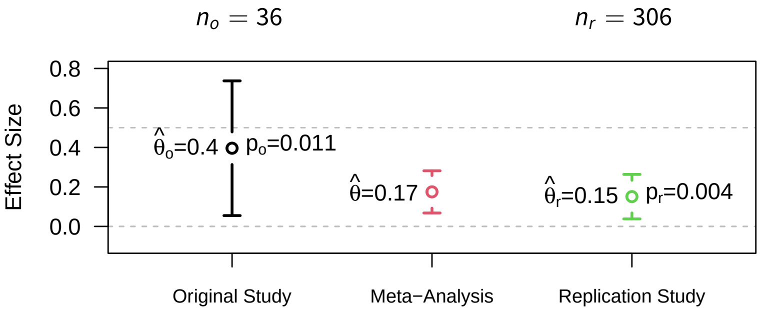
Assessement of Replication Success

- Limitations of currently used methods:
  - Two-trials rule is based on "double dichotomisation"
  - Q-test provides no information about true effect
  - Meta-analysis assumes exchangeability
- I will describe two reverse-Bayes approaches
  - without "double dichotomisation"
  - without exchangeability assumptions
  - but with explicit penalisation of effect size shrinkage

1. The sceptical p-value
2. The sceptical Bayes factor

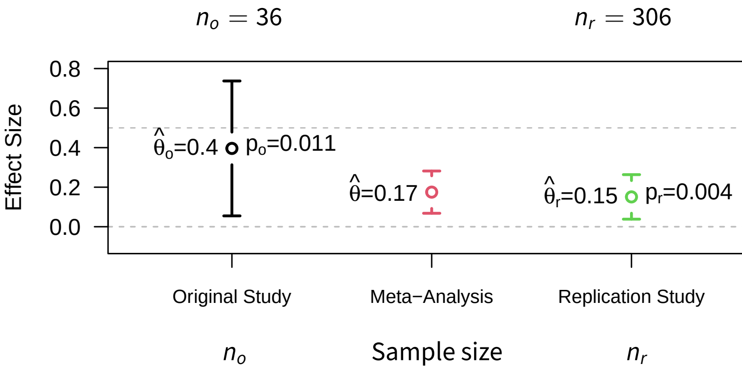
Example: Pyc and Rawson (2010), Social Sciences

Effect estimates with 95% confidence interval



Example: Pyc and Rawson (2010), Social Sciences

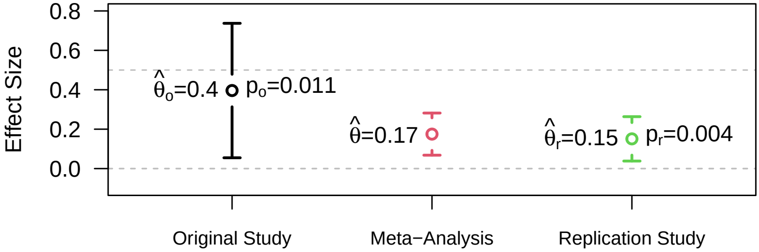
Effect estimates with 95% confidence interval



Example: Pyc and Rawson (2010), Social Sciences

Effect estimates with 95% confidence interval

$n_o = 36$ $n_r = 306$

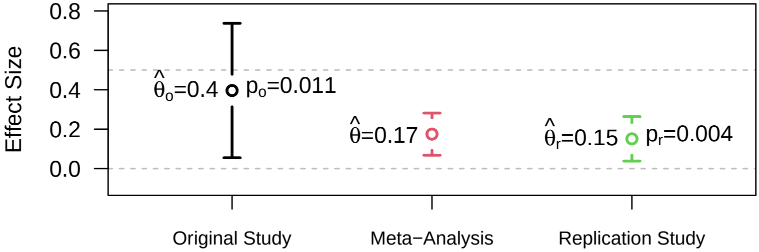


n_o	Sample size	n_r
$\hat{\theta}_o$	Effect estimate	$\hat{\theta}_r$

Example: Pyc and Rawson (2010), Social Sciences

Effect estimates with 95% confidence interval

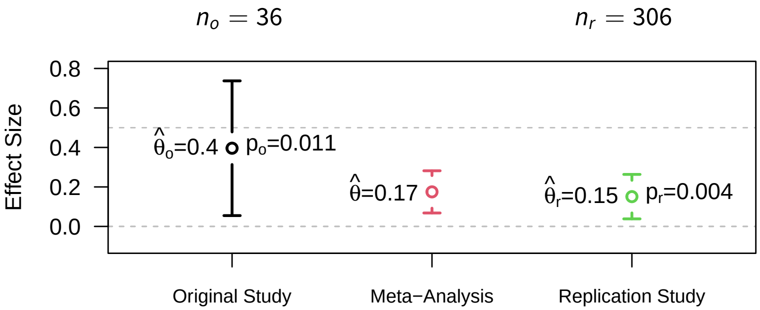
$n_o = 36$ $\qquad$ $n_r = 306$



n_o	Sample size	n_r
$\hat{\theta}_o$	Effect estimate	$\hat{\theta}_r$
σ_o	Standard error	σ_r

Example: Pyc and Rawson (2010), Social Sciences

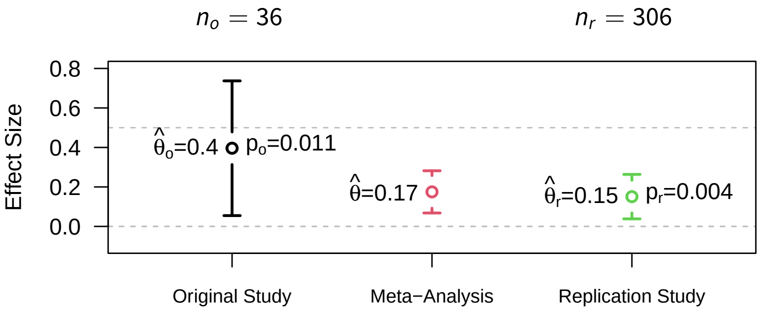
Effect estimates with 95% confidence interval



n_o	Sample size	n_r
$\hat{\theta}_o$	Effect estimate	$\hat{\theta}_r$
σ_o	Standard error	σ_r
z_o	z-value	z_r

Example: Pyc and Rawson (2010), Social Sciences

Effect estimates with 95% confidence interval



n_o	Sample size	n_r
$\hat{\theta}_o$	Effect estimate	$\hat{\theta}_r$
σ_o	Standard error	σ_r
z_o	z-value	z_r
p_o	one-sided p-value	p_r

A New Approach to Define Replication Success

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)

A new standard for the analysis and design of replication studies

Leonhard Held

University of Zurich, Switzerland

[*Read before* The Royal Statistical Society *at a meeting on 'Signs and sizes: understanding and replicating statistical findings' at the Society's 2019 annual conference in Belfast on Wednesday, September 4th, 2019, the President,* Professor D. Ashby, *in the Chair*]

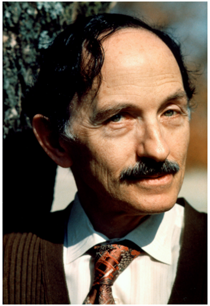
– A Bayes/non-Bayes compromise based on
  1. Reverse-Bayes analysis
  2. Quantification of prior-data conflict

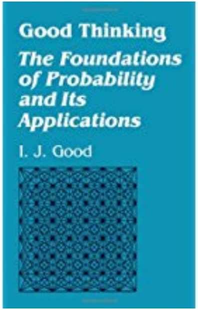
$\rightarrow$ The sceptical *p*-value p_S quantifies degree of replication success

Reverse-Bayes Analysis

Jack Good (1916-2009)

"We can make judgments of initial probabilities and infer final ones, or we can equally make judgments of final ones and infer initial ones by **Bayes's theorem in reverse***."*





Forward- and Reverse-Bayes

Bayesian updating

$f(\theta)$ $\quad f(\text{data} \mid \theta)$ $\quad f(\theta \mid \text{data})$

Forward-Bayes
inference

Forward- and Reverse-Bayes

Bayesian updating

$$f(\theta) \qquad f(\text{data} \mid \theta) \quad f(\theta \mid \text{data})$$

Bayesian downdating

Reverse-Bayes inference

Forward-Bayes inference

Forward- and Reverse-Bayes

Bayesian updating

$$f(\theta) \qquad f(\text{data} \mid \theta) \quad f(\theta \mid \text{data})$$

Bayesian downdating

Reverse-Bayes inference

Forward-Bayes inference

REVIEW

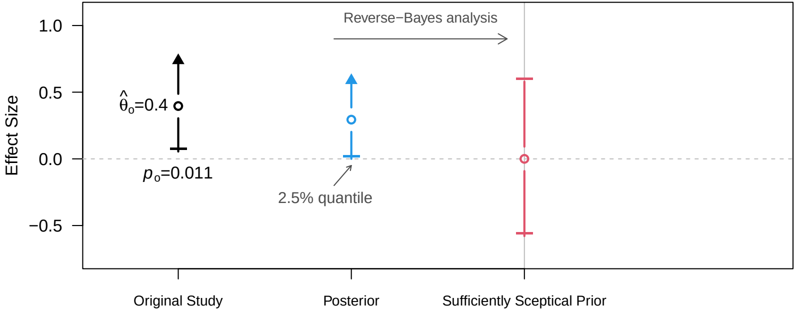
Research Synthesis Methods WILEY

Reverse-Bayes methods for evidence assessment and research synthesis

Leonhard Held[1] | Robert Matthews[2] | Manuela Ott[1,3] | Samuel Pawel[1]

The Proposed Approach: Step 1

One-sided $\alpha = 2.5\%$

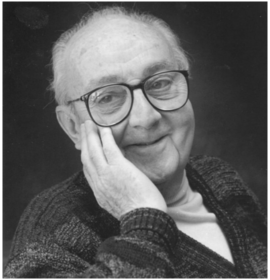


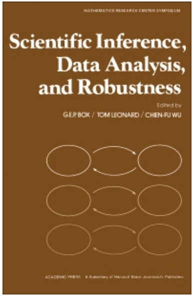
– Determine the variance τ^2 of a sceptical prior $\mathrm{N}(0, \tau^2)$ that makes the original result no longer convincing.

Prior-Data Conflict

George Box (1919-2013)

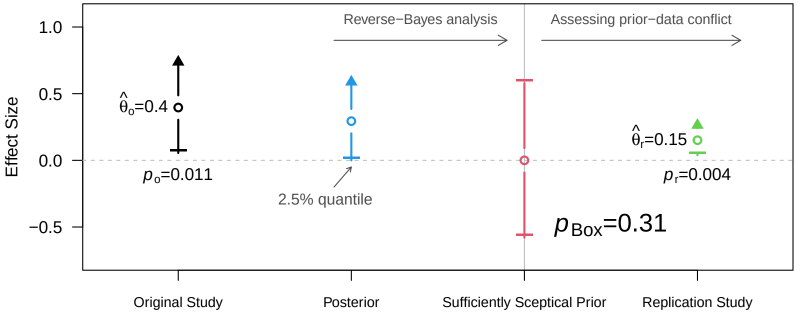
"The process of scientific investigation involves not one but two kinds of inference: estimation and criticism, used iteratively and in alternation."





The Proposed Approach: Step 2

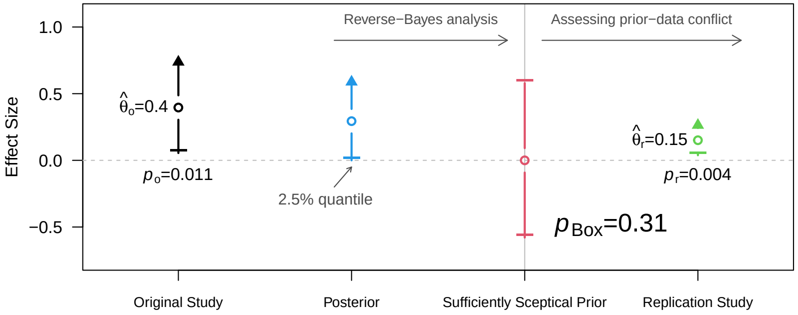
One-sided $\alpha = 2.5\%$



– Prior-data conflict is quantified based on the tail probability of the prior-predictive distribution: $p_{\text{Box}} = \Pr\{\mathrm{N}(0, \tau^2 + \sigma_r^2) \geq \hat{\theta}_r\}$.

The Proposed Approach: Step 2

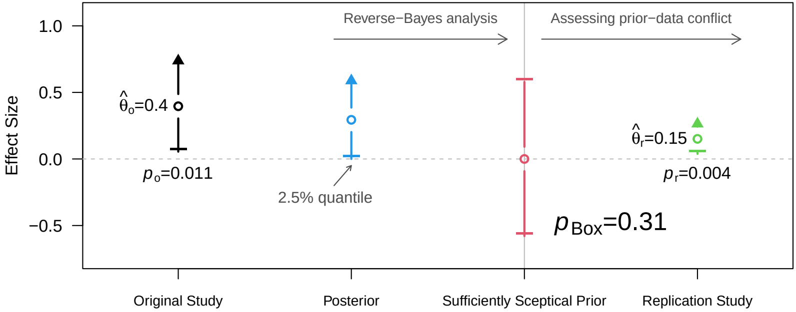
One-sided $\alpha = 2.5\%$



- Prior-data conflict is quantified based on the tail probability of the prior-predictive distribution: $p_{\text{Box}} = \Pr\{\text{N}(0, \tau^2 + \sigma_r^2) \geq \hat{\theta}_r\}$.
- Conflict between the sceptical prior and the replication effect estimate ($p_{\text{Box}} \leq \alpha$) defines replication success at level α.

The Proposed Approach

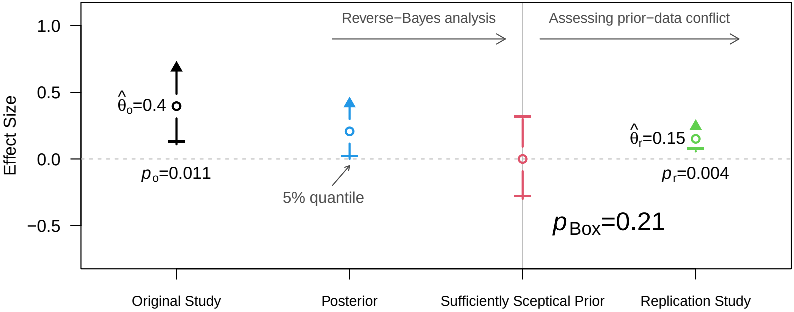
One-sided $\alpha = 2.5\%$



No replication success at level $\alpha = 2.5\%$

The Proposed Approach

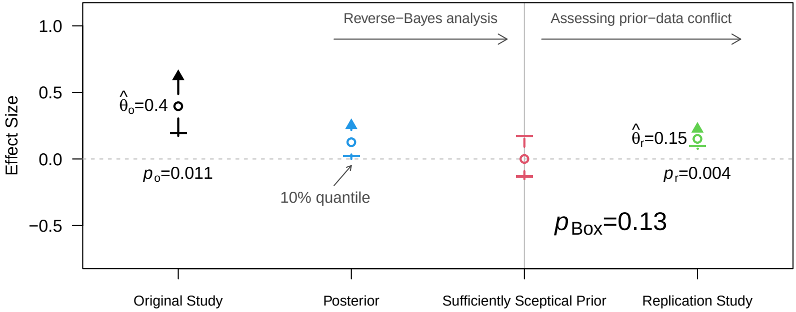
One-sided $\alpha = 5\%$



No replication success at level $\alpha = 5\%$

The Proposed Approach

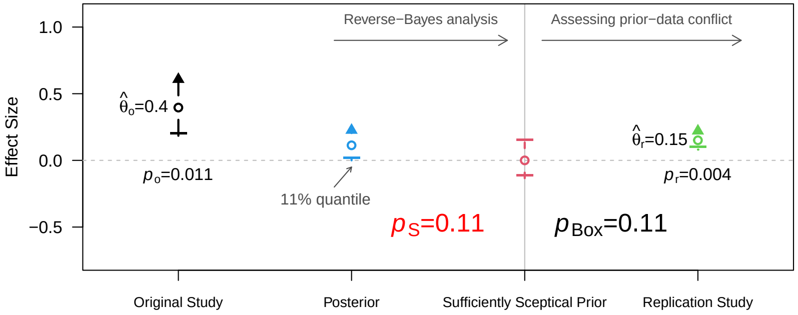
One-sided $\alpha = 10\%$



No replication success at level $\alpha = 10\%$

The Proposed Approach

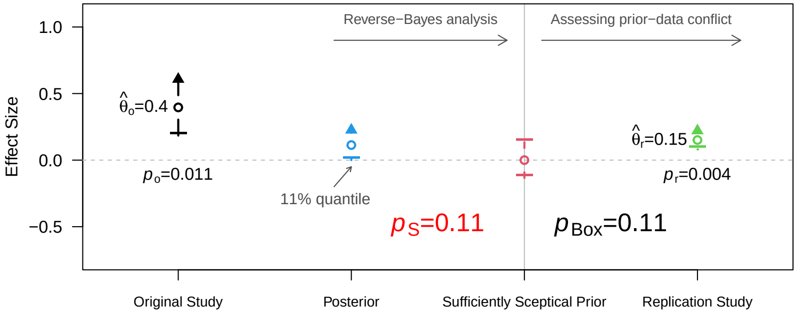
One-sided $\alpha = 11\%$



Replication success at level $\alpha = 11\%$

The Proposed Approach

One-sided $\alpha = 11\%$



Replication success at level $\alpha = 11\%$

The smallest level α where $p_{\text{Box}} \leq \alpha$ is the sceptical p-value p_S

The Sceptical p-Value

- always exists, fulfills $p_S > \max\{p_o, p_r\}$

The Sceptical p-Value

- always exists, fulfills $p_S > \max\{p_o, p_r\}$
- does not depend on α

The Sceptical p-Value

- always exists, fulfills $p_S > \max\{p_o, p_r\}$
- does not depend on α
- can be computed analytically under standard normality assumptions

The Sceptical p-Value

- always exists, fulfills $p_S > \max\{p_o, p_r\}$
- does not depend on α
- can be computed analytically under standard normality assumptions
- depends on both z-values z_o and z_r (resp. p-values p_o and p_r) and the relative sample size $c = n_r/n_o$:

$$
\begin{aligned}
p_S &= 1 - \Phi(|z_S|) \quad \text{where} \\
z_S^2 &= \begin{cases} z_H^2/2 & \text{for } c = 1 \\ \frac{z_A^2}{c-1}\left\{\sqrt{1+(c-1)z_H^2/z_A^2} - 1\right\} & \text{for } c \neq 1 \end{cases}
\end{aligned}
$$

where z_A^2 and z_H^2 is the arithmetic resp. harmonic mean of z_o^2 and z_r^2.

Replication Success in Terms of Relative Effect Size

Goal: Comparison of

- **sceptical *p*-value**
- **two-trials rule**
- **meta-analysis**

Replication Success in Terms of Relative Effect Size

Goal: Comparison of

- **sceptical *p*-value**
- **two-trials rule**
- **meta-analysis**

Key: Formulation in terms of

- **original *p*-value** p_o
- **relative effect size** $d = \hat{\theta}_r / \hat{\theta}_o$
- **relative sample size** $c = n_r / n_o$

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 706–720
https://doi.org/10.1214/21-AOAS1502

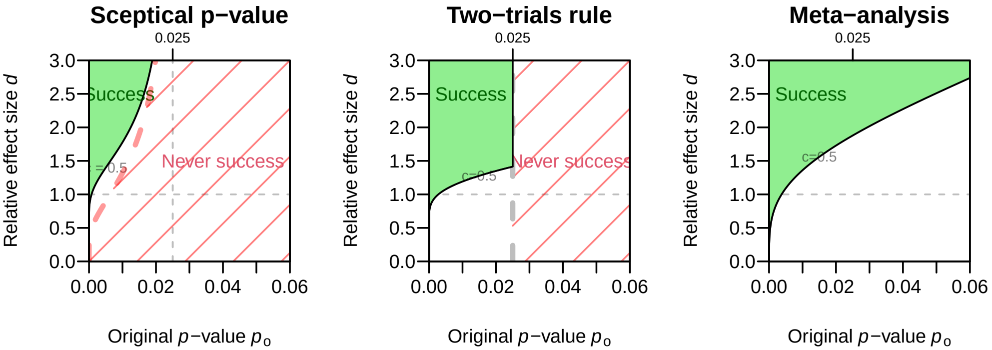
THE ASSESSMENT OF REPLICATION SUCCESS BASED ON RELATIVE EFFECT SIZE

BY LEONHARD HELD[a], CHARLOTTE MICHELOUD[b] AND SAMUEL PAWEL[c]

Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich,
[a]*leonhard.held@uzh.ch*, [b]*charlotte.micheloud@uzh.ch*, [c]*samuel.pawel@uzh.ch*

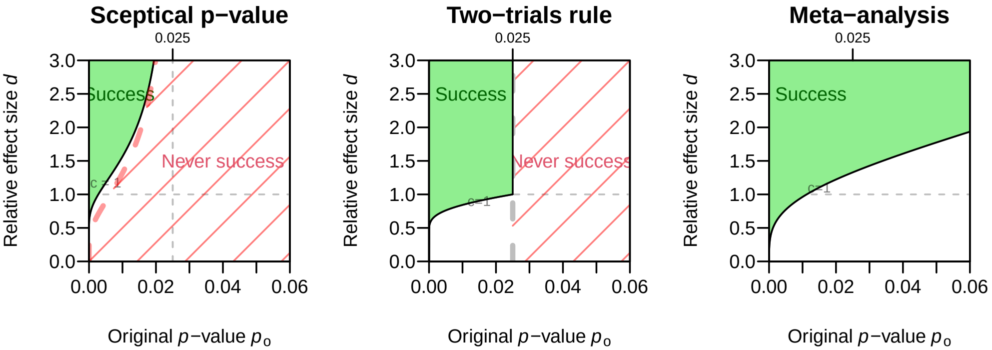
Replication Success Regions

Relative sample size $c = 0.5$



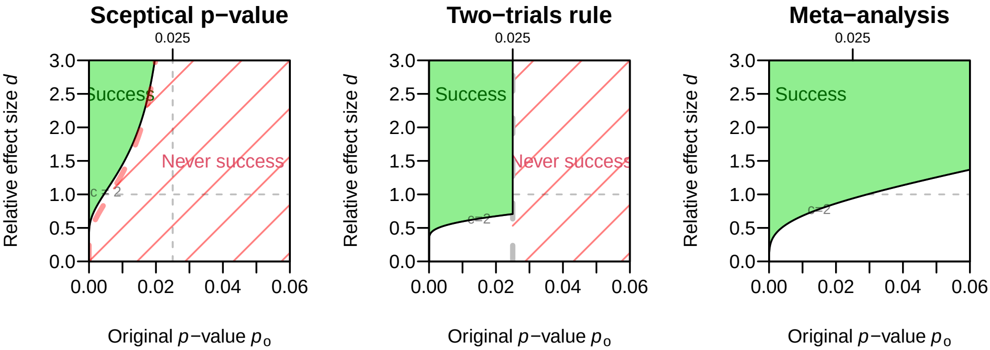
Replication Success Regions

Relative sample size $c = 1$



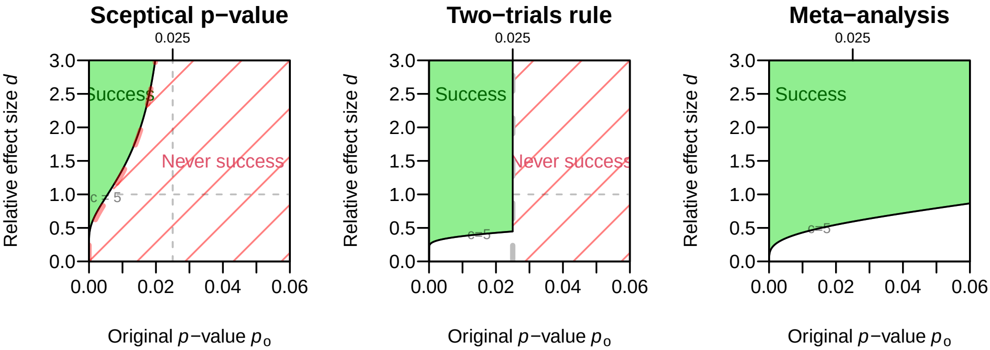
Replication Success Regions

Relative sample size $c = 2$



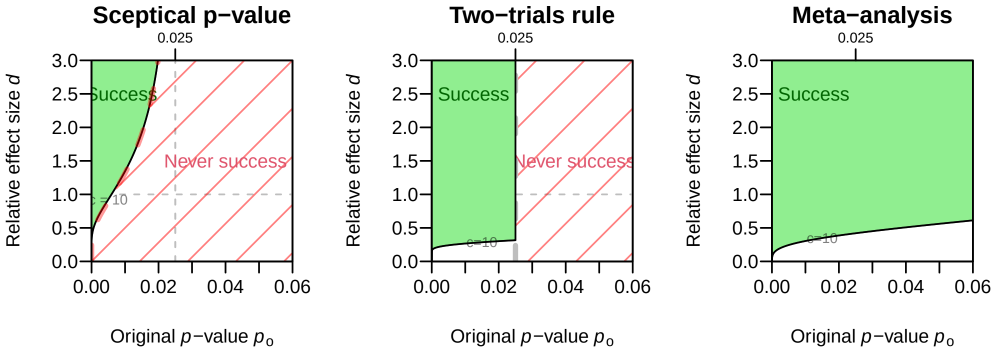
Replication Success Regions

Relative sample size $c = 5$



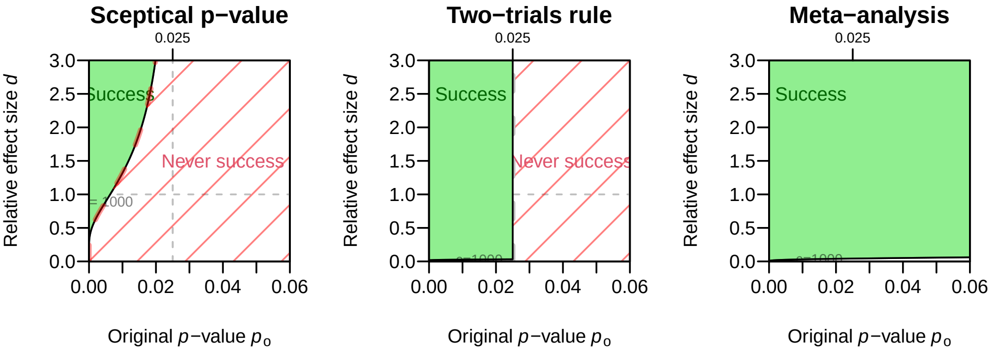
Replication Success Regions

Relative sample size $c = 10$



Replication Success Regions

Relative sample size $c = 1000$



Recalibration

Problem:

Nominal sceptical p-value is too stringent: Replication success is impossible for borderline significant original studies ($p_o \approx \alpha$).

Recalibration

Problem:

Nominal sceptical p-value is too stringent: Replication success is impossible for borderline significant original studies ($p_o \approx \alpha$).

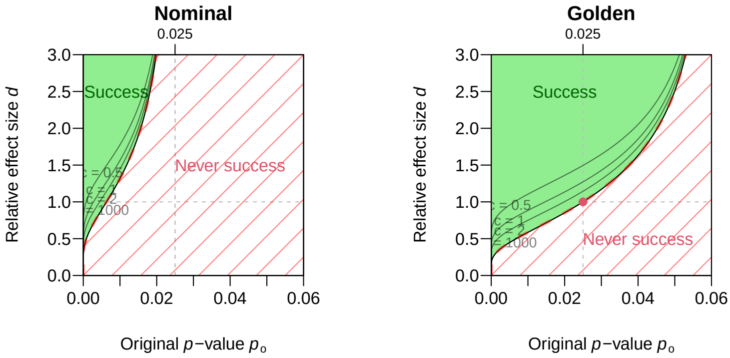
Solution:

Golden recalibration to

$$
\begin{aligned}
p_S &= 1 - \Phi(\sqrt{\varphi}\,|z_S|) \\
\text{where } \varphi &= (\sqrt{5}+1)/2 \approx 1.62
\end{aligned}
$$

is the golden ratio.

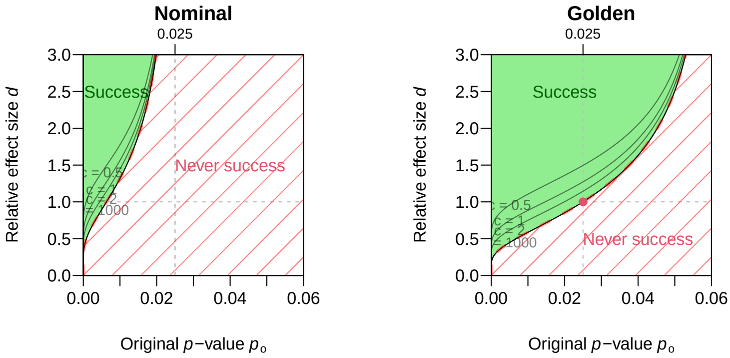
Nominal vs. Golden Sceptical P-Value



For a borderline convincing original result ($p_o \approx 0.025$), replication success

– is impossible with nominal p_S

Nominal vs. Golden Sceptical *P*-Value



For a borderline convincing original result ($p_o \approx 0.025$), replication success

- is impossible with nominal p_S
- is possible with golden p_S, if there is no effect size shrinkage.

Replication Projects
Proportion of successful replications

Project	Sample size	Two-trials rule (%)	Sceptical p-value (%)
Psychology	73	28.8	30.1
Social Sciences	21	61.9	52.4
Experimental Philosophy	31	74.2	71.0
Experimental Economics	18	55.6	55.6

Proportion of successful replications with the two-trials rule and the golden sceptical *p*-value ($\alpha = 2.5\%$)

When Do They Disagree?

Study	Project	c	d	p_o	p_r	p_S
Schmidt and Besner (2008)	Psychology	2.58	1.28	**0.028**	< 0.0001	0.024
Oberauer (2008)	Psychology	0.60	0.67	0.0003	**0.035**	0.017
Payne et al. (2008)	Psychology	2.65	0.41	0.001	0.023	**0.031**
Balafoutas and Sutter (2012)	Social Sciences	3.48	0.52	0.009	0.011	**0.04**
Pyc and Rawson (2010)	Social Sciences	9.18	0.38	0.011	0.004	**0.061**
Nichols (2006)	Experimental Philosophy	9.40	0.49	0.015	0.0006	**0.049**

p_S: golden sceptical p-value

When Do They Disagree?

Study	Project	c	d	p_o	p_r	p_S
Schmidt and Besner (2008)	Psychology	2.58	1.28	**0.028**	< 0.0001	0.024
Oberauer (2008)	Psychology	0.60	0.67	0.0003	**0.035**	0.017
Payne et al. (2008)	Psychology	2.65	0.41	0.001	0.023	**0.031**
Balafoutas and Sutter (2012)	Social Sciences	3.48	0.52	0.009	0.011	**0.04**
Pyc and Rawson (2010)	Social Sciences	9.18	0.38	0.011	0.004	**0.061**
Nichols (2006)	Experimental Philosophy	9.40	0.49	0.015	0.0006	**0.049**

p_S: golden sceptical p-value

Sceptical p-value

- does not require both studies to be significant

When Do They Disagree?

Study	Project	c	d	p_o	p_r	p_S
Schmidt and Besner (2008)	Psychology	2.58	1.28	**0.028**	< 0.0001	0.024
Oberauer (2008)	Psychology	0.60	0.67	0.0003	**0.035**	0.017
Payne et al. (2008)	Psychology	2.65	0.41	0.001	0.023	**0.031**
Balafoutas and Sutter (2012)	Social Sciences	3.48	0.52	0.009	0.011	**0.04**
Pyc and Rawson (2010)	Social Sciences	9.18	0.38	0.011	0.004	**0.061**
Nichols (2006)	Experimental Philosophy	9.40	0.49	0.015	0.0006	**0.049**

p_S: golden sceptical p-value

Sceptical p-value

- does not require both studies to be significant
- penalizes shrinkage

How Best to Quantify Replication Success?

ROYAL SOCIETY
OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Muradchanian J, Hoekstra R, Kiers H, van Ravenzwaaij D. 2021 How best to quantify replication success? A simulation study on the comparison of replication success metrics. *R. Soc. Open Sci.* **8**: 201697. https://doi.org/10.1098/rsos.201697

How best to quantify replication success? A simulation study on the comparison of replication success metrics

Jasmine Muradchanian, Rink Hoekstra, Henk Kiers and Don van Ravenzwaaij

Behavioural and Social Sciences, University of Groningen, The Netherlands

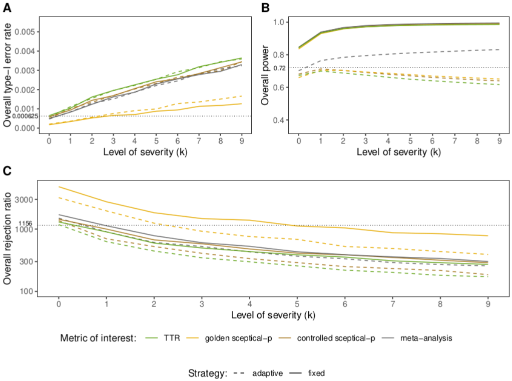
“The sceptical p-value performed particularly well under scenarios of high publication bias.”

Replication Success under Questionable Research Practices

Replication success under questionable research practices – a simulation study

Francesca Freuli* Leonhard Held† Rachel Heyard‡

https://osf.io/preprints/metaarxiv/s4b65/



Application to Social Sciences Replication Project

Study	$\hat{\theta}_r / \hat{\theta}_o$	n_r / n_o	p_o	p_r	p_S	$\tilde{p}_S$
Hauser et al. (2014), Nature	1.00	0.50	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Aviezer et al. (2012), Science	0.60	0.90	< 0.0001	< 0.0001	0.0003	< 0.0001
Wilson et al. (2014), Science	0.80	1.30	< 0.0001	< 0.0001	0.002	0.0001
Derex et al. (2013), Nature	0.60	1.30	< 0.0001	0.001	0.01	0.002
Karpicke and Blunt (2011), Science	0.60	1.20	< 0.0001	0.003	0.012	0.002
Janssen et al. (2010), Science	0.50	0.60	< 0.0001	0.013	0.017	0.003
Gneezy et al. (2014), Science	0.80	2.30	0.001	0.0001	0.019	0.004
Kovacs et al. (2010), Science	1.40	4.40	0.013	< 0.0001	0.03	0.009
Morewedge et al. (2010), Science	0.80	3.00	0.004	0.0003	0.036	0.011
Duncan et al. (2012), Science	0.60	7.40	0.002	< 0.0001	0.036	0.011
Nishi et al. (2015), Nature	0.60	2.40	0.002	0.005	0.046	0.016
Balafoutas and Sutter (2012), Science	0.50	3.50	0.009	0.011	0.085	0.04
Pyc and Rawson (2010), Science	0.40	9.20	0.011	0.004	0.11	0.061
Rand et al. (2012), Nature	0.20	6.30	0.004	0.12	0.19	0.13
Ackerman et al. (2010), Science	0.20	11.70	0.024	0.063	0.21	0.15
Sparrow et al. (2011), Science	0.10	3.50	0.0009	0.23	0.24	0.19
Shah et al. (2012), Science	-0.10	11.60	0.023	0.65	0.63	0.66
Kidd and Castano (2013), Science	-0.10	8.60	0.006	0.77	0.72	0.77
Gervais and Norenzayan (2012), Science	-0.10	9.80	0.014	0.79	0.73	0.78
Lee and Schwarz (2010), Science	-0.10	7.60	0.006	0.78	0.74	0.79
Ramirez and Beilock (2011), Science	-0.10	4.50	< 0.0001	0.80	0.79	0.85

Introduction

The Sceptical *p*-Value

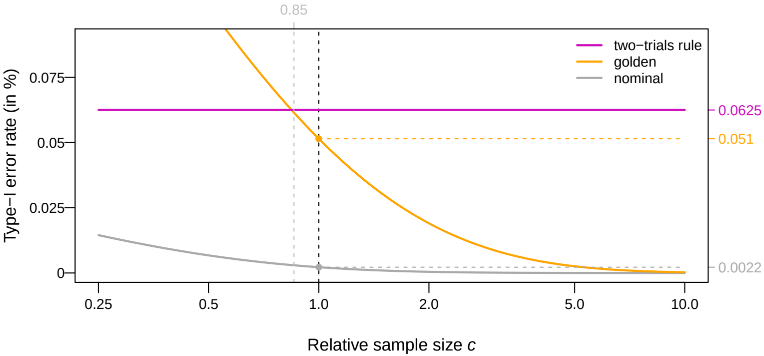
Type-I Error Control

The Sceptical Bayes Factor

Discussion and Epilogue

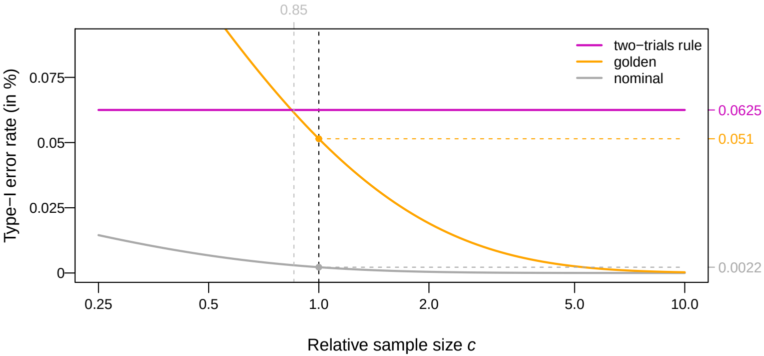
Overall Type-I Error Rate

Success probability over both studies under the null hypothesis



Overall Type-I Error Rate

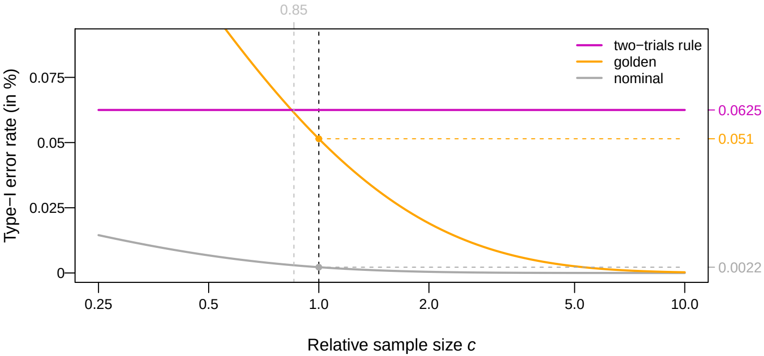
Success probability over both studies under the null hypothesis



Can we achieve exact overall T1E control for all values of c?

Overall Type-I Error Rate

Success probability over both studies under the null hypothesis



Can we achieve exact overall T1E control for all values of c?

$\rightarrow$ Controlled sceptical p-value

The Harmonic Mean χ^2 Test: $c = 1$

Suppose $z_o^2, z_r^2 \overset{\text{iid}}{\sim} \chi^2(1)$. We need the null distribution of

$$z_S^2 = z_H^2/2 = 1/(1/z_o^2 + 1/z_r^2)$$

$\rightarrow$ z_S^2 has a $\text{Ga}(1/2, 2)$ null distribution with cdf $F_1(.)$

$\rightarrow$ $p = 1 - F_1(z_S^2)$ has exact T1E control.

Appl. Statist. (2020)

The harmonic mean χ^2-test to substantiate scientific findings

Leonhard Held

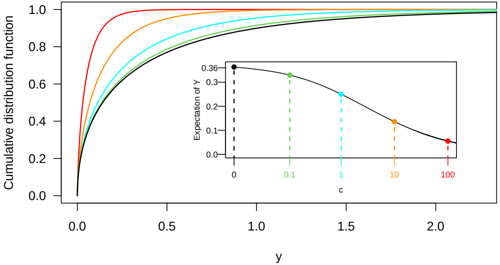
University of Zurich, Switzerland

The Case $c \neq 1$

Null distribution of $z_S^2 = \frac{z_A^2}{c-1}\left\{\sqrt{1+(c-1)z_H^2/z_A^2}-1\right\}$ required

– z_A^2 and z_H^2 are dependent, but z_A^2 and z_H^2/z_A^2 are independent

$\rightarrow$ cdf $F_c(.)$ of z_S^2 is available with one-dimensional numerical integration:



$\rightarrow$ $p = 1 - F_c(z_S^2)$ has exact T1E control.

A New Family of Combination Tests

Type I error control at $\alpha^2 = 0.025^2$

A Statistical Framework for Replicability

Leonhard Held*, Charlotte Micheloud* and Fadoua Balabdaoui†

*University of Zurich

Epidemiology, Biostatistics and Prevention Institute (EBPI)

and Center for Reproducible Science (CRS)

Hirschengraben 84, 8001 Zurich, Switzerland

and

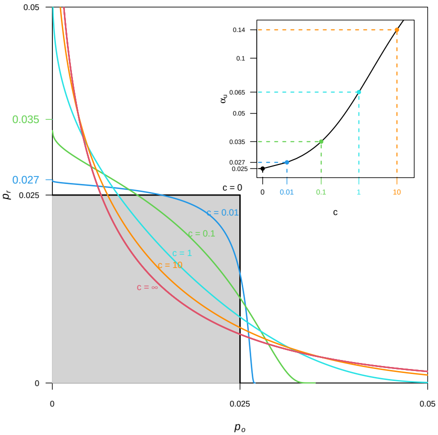
†ETH Zurich

Seminar für Statistik

Rämistrasse 101, 8092 Zürich, Switzerland

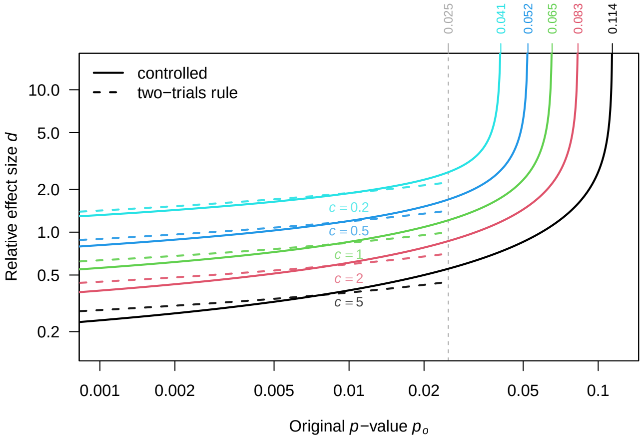
4th July 2022

https://arxiv.org/abs/2207.00464



Minimum Relative Effect Size

Threshold for replication success on relative effect size $d = \hat{\theta}_r / \hat{\theta}_o$



P-value Function and Confidence Interval

- Consider the generalized z-statistic

$$z_i(\mu) = \frac{\hat{\theta}_i - \mu}{\sigma_i} \qquad i \in \{o, r\}$$

  for the null hypothesis H_0: $\theta = \mu$.
- The z-values $z_o(\mu)$ and $z_r(\mu)$ are now used to compute $z_S^2(\mu)$.

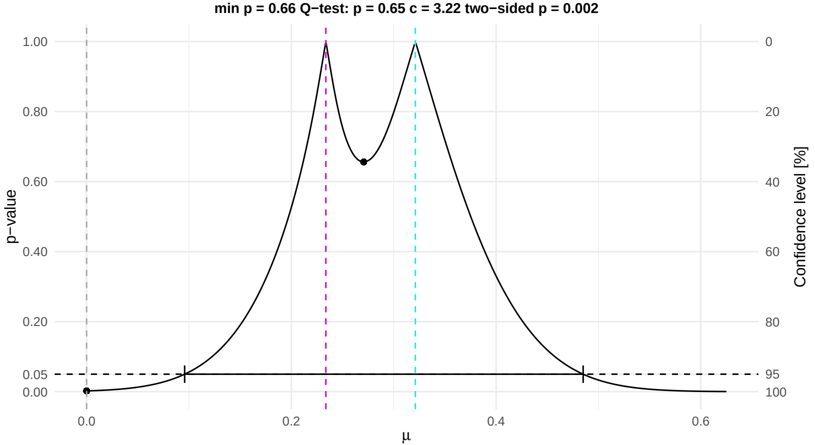
$\to$ A p-value function can be computed:

$$p(\mu) = 1 - F_c(z_S^2(\mu))$$

- Exact confidence intervals can be derived.

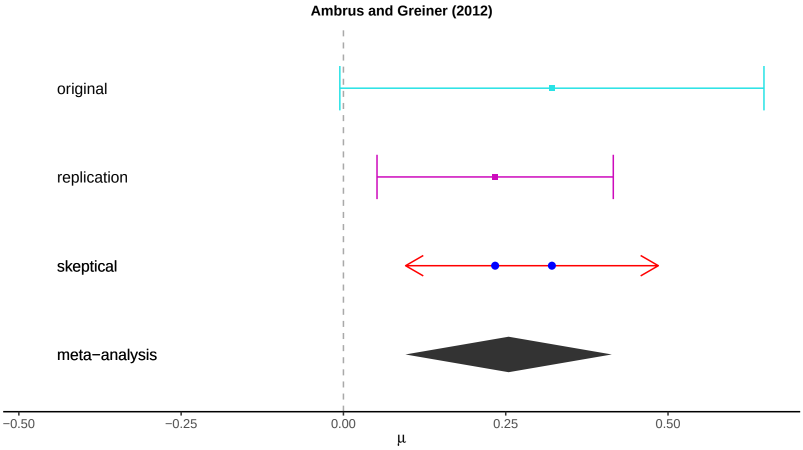
Ambrus and Greiner (2012)

$c = 3.22$, one-sided $p_S = 0.024$



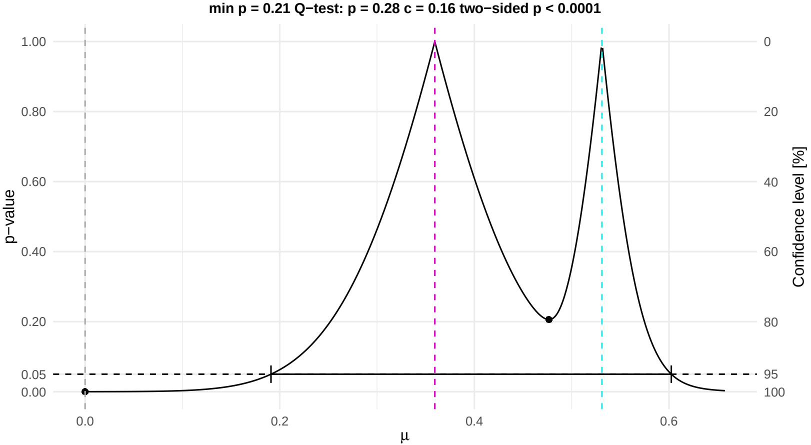
Ambrus and Greiner (2012)

Forest plot



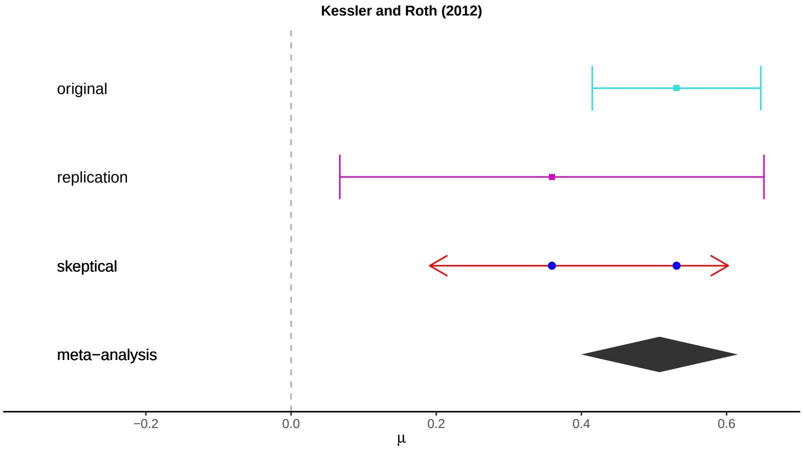
Kessler and Roth (2012)

$c = 0.16$, one-sided $p_S = 0.003$



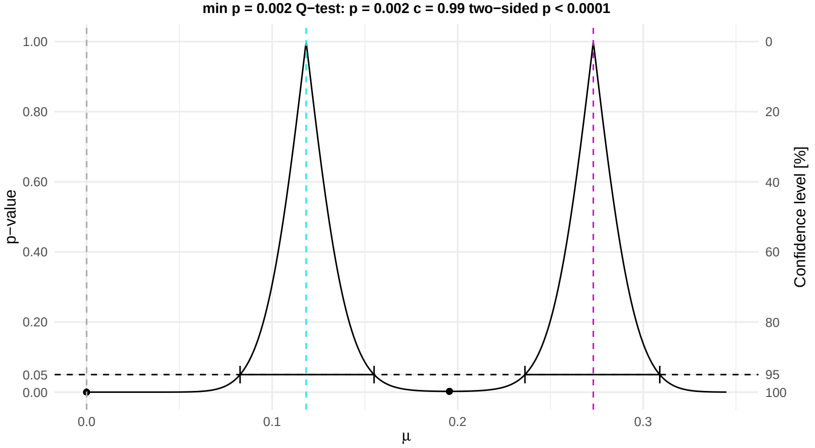
Kessler and Roth (2012)

Forest plot



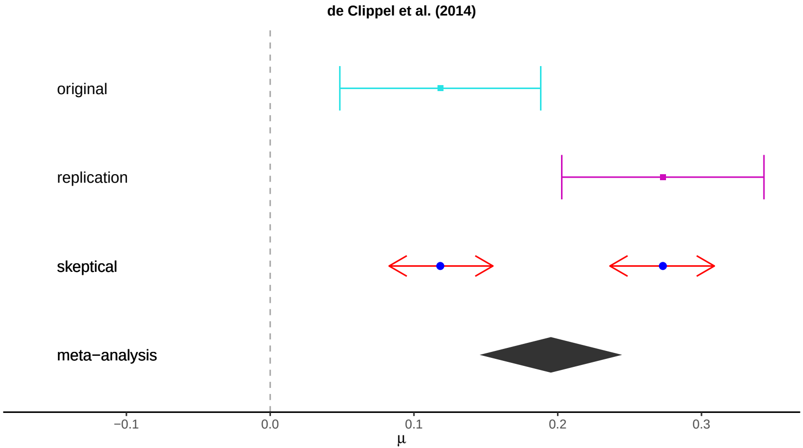
de Clippel et al. (2014)

$c = 0.99$, one-sided $p_S < 0.0001$



de Clippel et al. (2014)

Forest plot



Reverse-Bayes Assessment of Replication Studies with Bayes Factors

DOI: 10.1111/rssb.12491

ORIGINAL ARTICLE

The sceptical Bayes factor for the assessment of replication success

Samuel Pawel | **Leonhard Held**

Main idea

Reverse-Bayes Assessment of Replication Studies with Bayes Factors

DOI: 10.1111/rssb.12491

ORIGINAL ARTICLE

The sceptical Bayes factor for the assessment of replication success

Samuel Pawel | **Leonhard Held**

Main idea

1. Determine sceptical prior so that the original finding is no longer convincing in terms of the Bayes factor (Pericchi, 2020; Consonni, 2019)

Reverse-Bayes Assessment of Replication Studies with Bayes Factors

DOI: 10.1111/rssb.12491

ORIGINAL ARTICLE

The sceptical Bayes factor for the assessment of replication success

Samuel Pawel | Leonhard Held

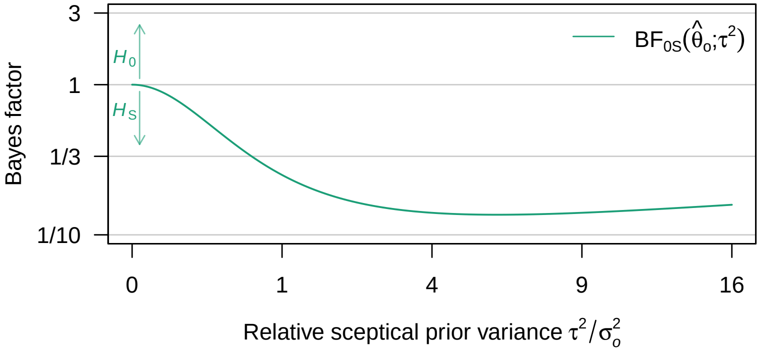
Main idea

1. Determine sceptical prior so that the original finding is no longer convincing in terms of the Bayes factor (Pericchi, 2020; Consonni, 2019)
2. Assess prior-data conflict of replication data and sceptical prior by contrasting it to an advocacy prior (posterior of effect size based on original study + flat prior) with another Bayes factor (Box, 1980)

Reverse-Bayes Assessment of Replication Studies with Bayes Factors

Bayes factor $\mathrm{BF}_{0S}(\hat{\theta}_o; \tau^2)$ for original data

$$H_0 : \theta = 0 \qquad \text{vs.} \qquad H_S : \theta \sim \mathrm{N}(0, \tau^2)$$



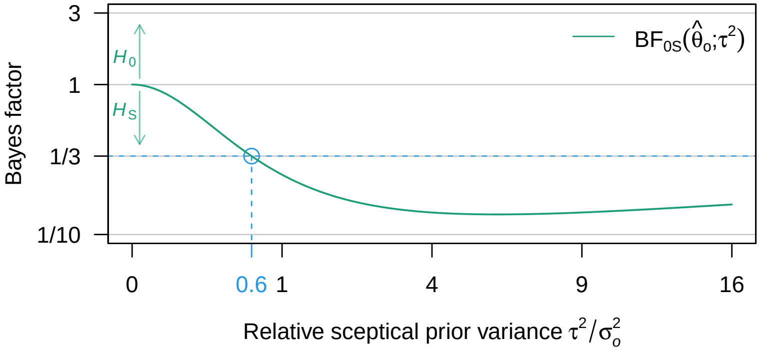
Reverse-Bayes Assessment of Replication Studies with Bayes Factors

Bayes factor $\mathrm{BF}_{0S}(\hat{\theta}_o; \tau^2)$ for original data

$$H_0 : \theta = 0 \qquad \text{vs.} \qquad H_S : \theta \sim \mathrm{N}(0, \tau^2)$$

$\rightarrow$ Reverse-Bayes step:

Choose $\tau^2 = \tau^2_\gamma$ such that evidence for H_0 is at level γ



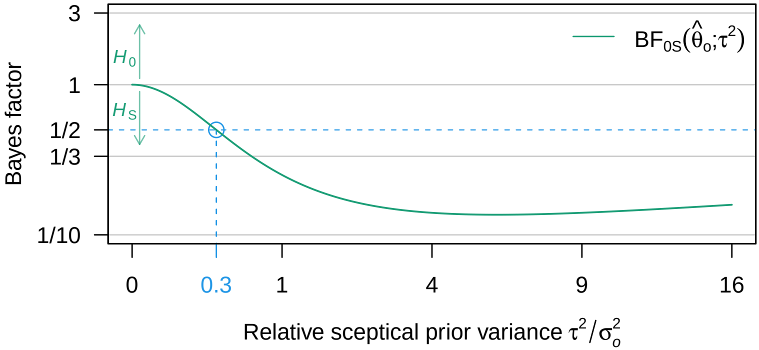
Reverse-Bayes Assessment of Replication Studies with Bayes Factors

Bayes factor $\mathrm{BF}_{0S}(\hat{\theta}_o; \tau^2)$ for original data

$$H_0: \theta = 0 \qquad \text{vs.} \qquad H_S: \theta \sim \mathrm{N}(0, \tau^2)$$

$\rightarrow$ Reverse-Bayes step:

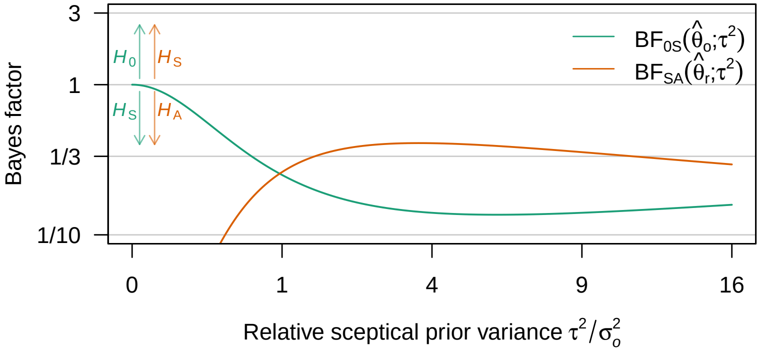
Choose $\tau^2 = \tau^2_\gamma$ such that evidence for H_0 is at level γ



Reverse-Bayes Assessment of Replication Studies with Bayes Factors

Bayes factor $\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; \tau^2)$ for replication data

$$H_{\mathrm{S}}: \theta \sim \mathrm{N}(0, \tau^2) \qquad \text{vs.} \qquad H_{\mathrm{A}}: \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$$



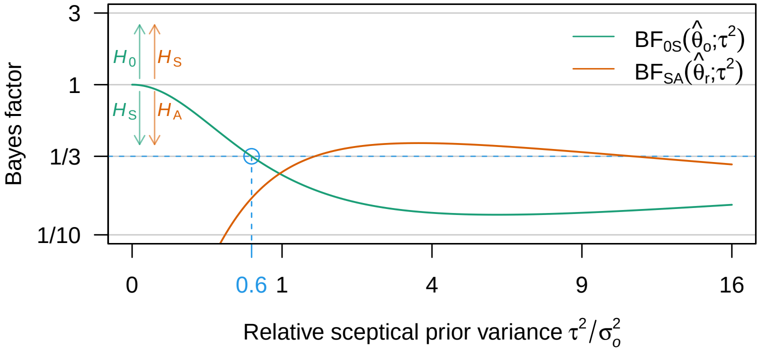
Reverse-Bayes Assessment of Replication Studies with Bayes Factors

Bayes factor $\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; \tau^2)$ for replication data

$$H_{\mathrm{S}}: \theta \sim \mathrm{N}(0, \tau^2) \qquad \text{vs.} \qquad H_{\mathrm{A}}: \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$$

$\rightarrow$ Replication success at level γ:

$$\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; \tau_\gamma^2) \leq \gamma$$



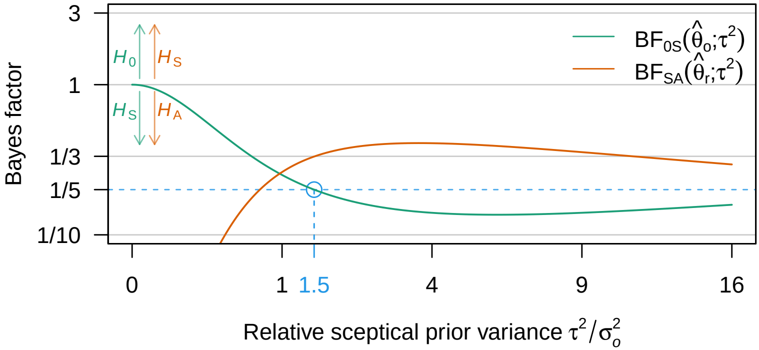
Reverse-Bayes Assessment of Replication Studies with Bayes Factors

Bayes factor $\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; \tau^2)$ for replication data

$$H_{\mathrm{S}}: \theta \sim \mathrm{N}(0, \tau^2) \qquad \text{vs.} \qquad H_{\mathrm{A}}: \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$$

$\rightarrow$ Replication success at level γ:

$$\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; \tau_\gamma^2) \leq \gamma$$

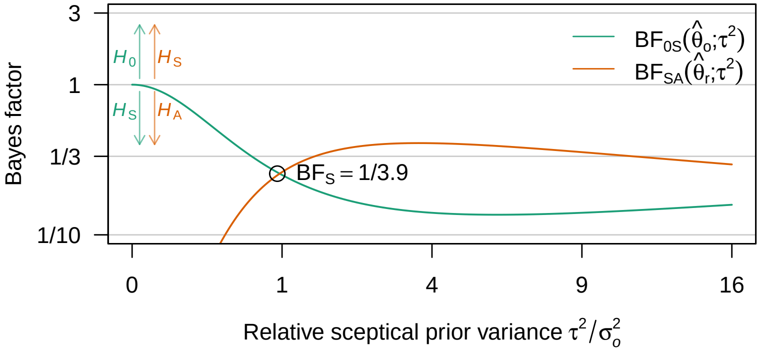


Reverse-Bayes Assessment of Replication Studies with Bayes Factors

Bayes factor $\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; \tau^2)$ for replication data

$$H_{\mathrm{S}}: \theta \sim \mathrm{N}(0, \tau^2) \qquad \text{vs.} \qquad H_{\mathrm{A}}: \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$$

$\rightarrow$ Sceptical Bayes factor BF_{S}: Smallest level γ at which $\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; \tau_\gamma^2) \leq \gamma$



The Sceptical Bayes Factor

Some properties

- Closed-form expression available when $\sigma_o = \sigma_r$ (involving Lambert W function)
- Cannot be smaller than the minimum Bayes factor from the original study
  $\rightarrow$ limited by the evidence from the original study
- BF_S depends on $Q = (\hat{\theta}_r - \hat{\theta}_o)^2/(\sigma_o^2 + \sigma_r^2)$ statistic
  $\rightarrow$ takes into account effect size compatibility
- Connected to the replication Bayes factor (Verhagen and Wagenmakers, 2014)
- BF_S may not exist

Application to Social Sciences Replication Project

Study	$\hat{\theta}_r / \hat{\theta}_o$	n_r / n_o	p_o	p_r	p_S	$\tilde{p}_S$	BF_S
Hauser et al. (2014), Nature	1.00	0.50	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 1/1000
Aviezer et al. (2012), Science	0.60	0.90	< 0.0001	< 0.0001	0.0003	< 0.0001	1/78
Wilson et al. (2014), Science	0.80	1.30	< 0.0001	< 0.0001	0.002	0.0001	1/45
Derex et al. (2013), Nature	0.60	1.30	< 0.0001	0.001	0.01	0.002	1/8.5
Karpicke and Blunt (2011), Science	0.60	1.20	< 0.0001	0.003	0.012	0.002	1/5.6
Janssen et al. (2010), Science	0.50	0.60	< 0.0001	0.013	0.017	0.003	1/1.6
Gneezy et al. (2014), Science	0.80	2.30	0.001	0.0001	0.019	0.004	1/6.9
Kovacs et al. (2010), Science	1.40	4.40	0.013	< 0.0001	0.03	0.009	1/3.2
Morewedge et al. (2010), Science	0.80	3.00	0.004	0.0003	0.036	0.011	1/3.9
Duncan et al. (2012), Science	0.60	7.40	0.002	< 0.0001	0.036	0.011	1/3.1
Nishi et al. (2015), Nature	0.60	2.40	0.002	0.005	0.046	0.016	1/2.5
Balafoutas and Sutter (2012), Science	0.50	3.50	0.009	0.011	0.085	0.04	1/1.6
Pyc and Rawson (2010), Science	0.40	9.20	0.011	0.004	0.11	0.061	1/1.2
Rand et al. (2012), Nature	0.20	6.30	0.004	0.12	0.19	0.13	
Ackerman et al. (2010), Science	0.20	11.70	0.024	0.063	0.21	0.15	
Sparrow et al. (2011), Science	0.10	3.50	0.0009	0.23	0.24	0.19	
Shah et al. (2012), Science	-0.10	11.60	0.023	0.65	0.63	0.66	
Kidd and Castano (2013), Science	-0.10	8.60	0.006	0.77	0.72	0.77	
Gervais and Norenzayan (2012), Science	-0.10	9.80	0.014	0.79	0.73	0.78	
Lee and Schwarz (2010), Science	-0.10	7.60	0.006	0.78	0.74	0.79	
Ramirez and Beilock (2011), Science	-0.10	4.50	< 0.0001	0.80	0.79	0.85	

Application to Social Sciences Replication Project

Study	$\hat{\theta}_r \,/\, \hat{\theta}_o$	$n_r \,/\, n_o$	p_o	p_r	p_S	$\tilde{p}_S$	BF_S
Hauser et al. (2014), Nature	1.00	0.50	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 1/1000
Aviezer et al. (2012), Science	0.60	0.90	< 0.0001	< 0.0001	0.0003	< 0.0001	1/78
Wilson et al. (2014), Science	0.80	1.30	< 0.0001	< 0.0001	0.002	0.0001	1/45
Derex et al. (2013), Nature	0.60	1.30	< 0.0001	0.001	0.01	0.002	1/8.5
Karpicke and Blunt (2011), Science	0.60	1.20	< 0.0001	0.003	0.012	0.002	1/5.6
Janssen et al. (2010), Science	0.50	0.60	< 0.0001	0.013	0.017	0.003	1/1.6
Gneezy et al. (2014), Science	0.80	2.30	0.001	0.0001	0.019	0.004	1/6.9
Kovacs et al. (2010), Science	1.40	4.40	0.013	< 0.0001	0.03	0.009	1/3.2
Morewedge et al. (2010), Science	0.80	3.00	0.004	0.0003	0.036	0.011	1/3.9
Duncan et al. (2012), Science	0.60	7.40	0.002	< 0.0001	0.036	0.011	1/3.1
Nishi et al. (2015), Nature	0.60	2.40	0.002	0.005	0.046	0.016	1/2.5
Balafoutas and Sutter (2012), Science	0.50	3.50	0.009	0.011	0.085	0.04	1/1.6
Pyc and Rawson (2010), Science	0.40	9.20	0.011	0.004	0.11	0.061	1/1.2
Rand et al. (2012), Nature	0.20	6.30	0.004	0.12	0.19	0.13	
Ackerman et al. (2010), Science	0.20	11.70	0.024	0.063	0.21	0.15	
Sparrow et al. (2011), Science	0.10	3.50	0.0009	0.23	0.24	0.19	
Shah et al. (2012), Science	-0.10	11.60	0.023	0.65	0.63	0.66	
Kidd and Castano (2013), Science	-0.10	8.60	0.006	0.77	0.72	0.77	
Gervais and Norenzayan (2012), Science	-0.10	9.80	0.014	0.79	0.73	0.78	
Lee and Schwarz (2010), Science	-0.10	7.60	0.006	0.78	0.74	0.79	
Ramirez and Beilock (2011), Science	-0.10	4.50	< 0.0001	0.80	0.79	0.85	

Janssen et al. (2010): $Q = 3.51 \rightarrow$ replication effect estimate is in conflict with advocacy prior

Discussion

Reverse-Bayes methods

- enable formalization of scepticism
- can be implemented with different measures of evidence
- require both studies to be convincing
- take into account effect size compatibility

Discussion

Reverse-Bayes methods

- enable formalization of scepticism
- can be implemented with different measures of evidence
- require both studies to be convincing
- take into account effect size compatibility

The methods

- can also be used for sample size calculations
- can include heterogeneity between studies
- can be extended to more than two replication studies

Epilogue: Are You a Bayesian?

"These days the statistician is often asked such questions as

'Are you a Bayesian?'

Epilogue: Are You a Bayesian?

“These days the statistician is often asked such questions as

’Are you a Bayesian?’

Epilogue: Are You a Bayesian?

“These days the statistician is often asked such questions as

> *’Are you a Bayesian?’*
> *’Are you a frequentist?’*

Epilogue: Are You a Bayesian?

“These days the statistician is often asked such questions as

> *’Are you a Bayesian?’*
> *’Are you a frequentist?’*
> *’Are you a data scientist?’*

Epilogue: Are You a Bayesian?

“These days the statistician is often asked such questions as

> *’Are you a Bayesian?’*
> *’Are you a frequentist?’*
> *’Are you a data scientist?’*
> *’Are you a designer of experiments?’*

Epilogue: Are You a Bayesian?

"These days the statistician is often asked such questions as

> *'Are you a Bayesian?'*
> *'Are you a frequentist?'*
> *'Are you a data scientist?'*
> *'Are you a designer of experiments?'*

I will argue that the appropriate answer to all these questions can be (and preferably should be) "yes", and that we can see why this is so if we consider the scientific context of what statisticians do."

Epilogue: Are You a Bayesian?

George Box (1983)

“These days the statistician is often asked such questions as

> *’Are you a Bayesian?’*
> *’Are you a frequentist?’*
> *’Are you a data analyst?’*
> *’Are you a designer of experiments?’*

I will argue that the appropriate answer to all these questions can be (and preferably should be) “yes”, and that we can see why this is so if we consider the scientific context of what statisticians do.”

An Apology for Ecumenism in Statistics

G. E. P. Box